Who Keeps the Recording?

What 53 Transcription and Meeting-Notes Services Say About Training, Retention and Access

Ankur Shrestha

Pepys, United States
ORCID: 0009-0009-6036-5849

13 November 2026

Abstract

People upload interviews, meetings and calls to AI transcription and meeting-notes services, and the public documents of those services are the main source a user, an ethics office or a procurement team can consult about what happens to the audio. We analyzed the privacy policies, terms of service, data processing agreements, sub-processor lists and AI-data pages of 53 services: 35 consumer and prosumer transcription tools, 10 meeting-notes bots, 5 quote-based services and 3 meeting platforms. All 247 documents were taken from Internet Archive captures, not live pages; 233 (94.3%) were retrieved. Two coders worked independently from the saved text under a twelve-variable codebook fixed before collection, every quote was checked as an exact substring of the saved document, and disagreements were adjudicated against the text. Only 14 services (26.4%, 95% CI 16.4–39.6) state that customer content is not used to train AI models. Another 15 (28.3%) have at least one document that permits training by default, including 7 whose own documents contradict each other; for 18 (34.0%) the documents are silent or unclear. Of the 15, 8 describe no way to opt out. Audio retention is open-ended ("as long as necessary") at 22 services (41.5%) and a fixed period at 14 (26.4%). Eleven services (20.8%) describe routine human access to customer content. Compared with captures from October 2024, 23 of 41 comparable services are coded differently on training, and in 16 of those the 2024 policy had said nothing about it. In a companion survey of 130 universities, the tools institutions approve or bar most often do not line up with their own training terms. Inter-coder agreement ranged from $\kappa = 0.69$ to $\kappa = 0.97$. The documents give users and institutions an incomplete and sometimes contradictory account of whether a recording is used to train models.

Keywords: AI transcription; privacy policies; terms of service; model training; data retention; meeting assistants; document analysis.

1 Introduction

AI transcription has moved from a specialist service to an everyday tool. A researcher uploads interview recordings; a journalist records a call; a manager lets a notes bot join a meeting. In each case the audio leaves the user's device and goes to a company whose public documents are, for most users and for most institutions, the only description of what will happen to it. Those documents matter in two ways. They are what users can read before uploading, and they are what university ethics offices, IT departments and procurement teams cite when they approve or bar a tool. A companion survey of 130 universities found that institutions name specific transcription and meeting-notes products as approved or prohibited, and that 62 of the 90 institutions with guidance mention vendor training, data retention or recording deletion [1].

The questions are no longer hypothetical. In August 2026 a US federal court allowed privacy claims to proceed against Otter.ai in a consolidated class action; the order recites allegations that the company retained conversation contents and voiceprints and used them to train its speech-recognition models [2]. Those are allegations at the pleading stage, not findings, but they show which clauses are at stake: training on customer content, retention, human access and biometric data.

Privacy policies have been studied at scale for mobile apps [3] and websites [4], and recent work has coded how six frontier AI developers describe training on users' chat data [5]. We found no comparable analysis of transcription and meeting-notes services, where the content is audio of people who often did not choose the service themselves. This paper asks five descriptive questions:

- **RQ1.** What share of services say customer audio or transcripts may be used to train or improve AI models by default, only with consent, only in de-identified form, or never?
- **RQ2.** How long do the documents say audio and transcripts are kept, and how quickly are they deleted after the user deletes them?
- **RQ3.** What do the documents disclose about human access, sub-processors, AI providers and data location?
- **RQ4.** How often do they mention voiceprints or biometric identifiers, and do business customers get different terms?
- **RQ5.** How did training and retention language change between October 2024 and October 2026?

The study describes what documents say. It does not test what any service does with data, and nothing here is a finding about a company's practices.

2 Method

Design and registration. This is a cross-sectional document analysis with a two-year look-back. The frame, codebook and analysis plan were written and hashed (SHA-256) before collection, on 3 October 2026. The study was not registered in a public registry; the plan files and their hash log are in the released dataset. The deviations from the plan are listed at the end of this section.

Sample. The unit is the service. The frame is the 50-service list of a companion pricing audit, built on 30 September 2026 before any coding for this study (35 consumer and prosumer transcription tools, 10 meeting-notes bots, 5 quote-based services), plus 3 meeting platforms (Zoom, Microsoft Teams and Google Meet), added because Zoom and Teams are the tools institutions most often approve [1]. The frame includes Pepys, the author's affiliation, coded by the same procedure from its public documents; Pepys is one of 53 services and is reported in the same tables as the others. No service was replaced. The frame favors English-language services with a self-serve web presence, so the results describe that market, not every transcription provider.

Documents. For each service we looked for five document types: privacy policy, terms of service, a publicly posted data processing agreement, a public sub-processor list, and any page specifically about how customer content is used for AI (help-center, trust-center or policy pages). Documents were located by web search; 247 were identified (60 privacy policies, 59 terms, 19 data processing agreements, 22 sub-processor lists and 73 AI-data pages among those retrieved). Platform services were limited to the documents governing their transcription or meeting-AI features.

Source rule. Every document was taken from the Internet Archive's Wayback Machine: the newest capture with HTTP status 200 not after 3 October 2026, fetched from the raw `id_` endpoint, with no request to the service's own site. We did this so that collection is reproducible and so that no service's terms on automated access were engaged. Each capture was saved with SHA-256

hashes of the raw bytes and extracted text. For the privacy policy we also took the newest capture not after 3 October 2024 and 3 October 2025. Of 247 documents, 233 (94.3%, 95% CI 90.7–96.6) were retrieved; 14 were not archived or not retrievable, including every document of one service whose domain is excluded from the archive. Fifteen retrieved captures contained almost no text (JavaScript shells), and 38 were older than 1 April 2026; both were recorded and the coders were told to treat shells as unavailable. Capture dates range from November 2022 to 3 October 2026 (median August 2026).

Variables. The codebook (Appendix A) has twelve variables: training on customer content by default; training by third-party AI providers; opt-out mechanism; audio retention; transcript retention; time to delete after the user deletes content or closes the account; human review; sub-processor disclosure; whether AI providers are named; data location; voiceprint or biometric language; and whether business customers get different training or retention terms. Training is coded `prohibited` when a document says customer content is not used to train models, `yes_default` when content may be used unless the user opts out or no opt-out is described, `opt_in` when only with the user's affirmative choice, and `deidentified_or_aggregated_only` when training is limited to de-identified, anonymized or aggregated data. Where a service's documents disagree, the code is `conflicting` and both passages are quoted. For drift (RQ5), the 2024 and 2025 privacy-policy captures were coded on training and audio retention only.

Coding and adjudication. Two coders (A and B) coded every service independently from the saved document text only, without the web or outside knowledge, and without access to each other's work. Every value other than "not stated" required a verbatim quote of 5 to 60 words, and every quote was machine-checked as an exact substring of the saved text (after folding whitespace, quotation-mark and dash variants). Coding was language-model-assisted under the written codebook and author supervision. The two codings disagreed on 103 values across main and drift blocks. An adjudicator resolved each against the saved documents, choosing A's value, B's value or a third value, with a quote and a one-sentence reason; the adjudicated record is final. After adjudication there were no unresolved disagreements and no failed quotes.

Analysis. Counts and shares with Wilson 95% intervals over all 53 services; tabulation by service group; Cohen's κ and raw agreement between coders before adjudication; and, for RQ5, the share of services whose training or retention code changed between captures. No inferential tests were planned and no service is ranked.

Deviations and limits on the record. (1) Discovery was first meant to check archive availability for each URL; the archive rate-limited parallel requests, so discovery used web search only and all archive lookups ran in a single queue. (2) The author's own service had never been archived; the author requested captures of its privacy and terms pages on 3 October 2026, before the cutoff. No other service's pages were submitted. (3) Agreement on data location compares the category (stated, customer choice, not stated) rather than the wording of region lists. (4) The deletion variable was clarified after the first group of ten services to cover deletion after account closure as well as after deleting content; adjudication applied the clarified rule to all services. (5) An exploratory split of biometric mentions into voice-specific and generic was added after coding and is labeled as such. (6) Section 3.7, which sets institutions' tool lists from the companion survey beside each service's training code, was added after analysis and is descriptive. (7) The plan said named vendors would receive a 14-day right of reply before publication; this was not done, and no service was contacted before publication. Other limits: the 2026 codes draw on all documents while the 2024 and 2025 codes draw on the privacy policy alone, so RQ5 compares like with like only for the privacy policy; enterprise contracts are not public; and some services' documents were captured well before the cutoff.

Table 1: What the documents say about training AI models on customer audio or transcripts (n = 53 services).

Training on customer content	Services	%	95% CI
Prohibited (not used for training)	14	26.4	16.4–39.6
Opt-in only	5	9.4	4.1–20.3
De-identified or aggregated only	4	7.5	3.0–17.9
Yes, by default	4	7.5	3.0–17.9
Documents conflict	7	13.2	6.5–24.8
Unclear	7	13.2	6.5–24.8
Not stated	11	20.8	12.0–33.5
No documents retrievable	1	1.9	0.3–9.9

3 Results

3.1 Training on customer content (RQ1)

Only 14 of 53 services (26.4%, 95% CI 16.4–39.6) have documents that say customer content is not used to train AI models, without a contradicting document (Table 1). Five (9.4%) describe training only with the user's affirmative choice. Fifteen (28.3%, 18.0–41.6) have at least one document that permits training by default: 4 without qualification, 4 limited to de-identified or aggregated data, and 7 whose documents contradict each other. For 18 (34.0%) the documents are silent (11) or unclear (7). Taken together, for 25 services (47.2%, 34.4–60.3) the documents do not give a clear answer.

The conflicts are between documents published by the same company. In each of the seven, at least one document permits training and another says content is not used. Rev's terms say customer content "may be used for continuous training of those models", while a security FAQ on its site opens: "No. Rev does not use customer data, including transcripts, captions, or any other uploaded content, to train AI models". Descript's privacy policy describes an opt-out ("You can opt out of having your Projects used to improve the Descript Service"), while a security page describes the setting as "disabled by default". Otter.ai's privacy policy describes "training our proprietary AI technology on de-identified audio recordings and on transcriptions (which may contain Personal Information)", while its terms and security page describe training on de-identified or aggregated data. In several cases the contradiction is between a newer policy and an older help page, or between consumer and business documents; we code what the documents say and do not decide which one governs.

Opt-out routes are scarce where training is the default. Of the 15 services whose documents permit default training, 8 describe no way to opt out, 4 describe a request to the company, 3 describe a self-serve setting, and 1 limits opt-out to enterprise customers. Documents of 19 services (35.8%) say that third-party AI providers may not train on customer content; 2 permit it and 28 (52.8%) say nothing.

3.2 Retention and deletion (RQ2)

Audio retention is most often open-ended. The documents of 22 services (41.5%) say content is kept as long as necessary for stated purposes, 10 (18.9%) until the user deletes it, and 14 (26.4%) give a fixed period, from deletion right after processing to three years. Four say nothing and 2 conflict. Transcript retention follows the same pattern: open-ended at 22 (41.5%), until the user deletes at 16 (30.2%), and a fixed period at 8 (15.1%). Twenty-nine services (54.7%) state how long deletion

Table 2: Most common value of each other variable (n = 53). Full distributions with confidence intervals are in the dataset.

Variable	Most common value	Services (%)
Third-party providers training	Not stated	28 (52.8)
Opt-out mechanism	None stated	27 (50.9)
Audio retention	As long as necessary	22 (41.5)
Transcript retention	As long as necessary	22 (41.5)
Deletion after request	Fixed period stated	29 (54.7)
Human review	Conditional only	24 (45.3)
Sub-processors	Categories only	26 (49.1)
AI providers	Named	22 (41.5)
Data location	Stated	44 (83.0)
Biometric or voiceprint	Not stated	30 (56.6)
Business-tier differences	None stated	36 (67.9)

takes after a user deletes content or closes an account, most often 30 days; 22 (41.5%) mention deletion without a time.

3.3 Access: people, sub-processors, AI providers, location (RQ3)

Eleven services (20.8%) describe routine human access to customer content, for example annotation, quality review or transcription by human contractors; 24 (45.3%) describe access only under conditions such as support requests, user consent or legal obligation; 3 say staff never access content; and 10 say nothing. Twenty-five services (47.2%) publish a named list of sub-processors and 26 (49.1%) describe them only by category. Twenty-two (41.5%) name the third-party AI or speech providers that process content; 11 describe them by category and 18 do not mention them. Forty-four (83.0%) state where data is stored or processed.

3.4 Biometrics and business tiers (RQ4)

Twenty services (37.7%) mention voiceprints, voice data as biometric data, or biometric identifiers: 13 together with consent or notice terms and 7 without. In an exploratory split, 14 of those mentions (26.4% of services) concern voice or speaker data specifically; the rest are generic lists of special categories of personal data. Two services disclaim collecting biometric data. Fifteen services (28.3%) state different training or retention terms for business or enterprise customers than for self-serve users, for example training opt-outs or custom retention available only on business plans.

3.5 Change since 2024 (RQ5)

Forty-one services had a readable privacy-policy capture from October 2024 that could be compared with their 2026 coding. Twenty-three of them (56.1%, 41.0–70.1) are coded differently on training. In 16 of the 23 the 2024 policy said nothing about training; by 2026 six of those state that content is not used, three describe opt-in, three limit training to de-identified data, two have conflicting documents, one is unclear and one permits default training. Three services moved from permitting default training to conflicting documents, and one from default training to opt-in. Against October 2025 captures, 22 of 47 comparable services are coded differently. Because the 2026 codes draw on all documents and the earlier codes on the privacy policy alone, these figures describe how much more the documents now say about training, not a precise count of policy changes.

Table 3: Agreement between Coder A and Coder B before adjudication (n = 53).

Variable	Raw agreement	Cohen's κ
Training on customer content	0.91	0.89
Third-party providers training	0.92	0.87
Opt-out mechanism	0.98	0.97
Audio retention	0.81	0.74
Transcript retention	0.77	0.69
Deletion after request	0.83	0.69
Human review	0.83	0.77
Sub-processors	0.87	0.76
AI providers named	0.94	0.91
Data location (category)	0.91	0.83
Biometric or voiceprint	0.91	0.84
Business-tier differences	0.91	0.80

Table 4: Institutions (of 130) that name each service as approved or prohibited in guidance for research transcription, and the code for training on customer content from the service's own documents (n = 12 services in both studies).

Service	Institutions approving	Institutions prohibiting	Training code in its own documents
Zoom (AI features)	41	1	Prohibited
Microsoft Teams	20	0	Documents conflict
Rev	6	3	Documents conflict
Happy Scribe	1	0	Opt-in only
Trint	2	0	Prohibited
TurboScribe	1	0	Prohibited
Otter.ai	5	20	Documents conflict
Read.ai	0	17	Unclear
Fireflies.ai	1	11	Prohibited
MeetGeek	0	2	Opt-in only
Fathom	0	2	De-identified or aggregated only
Sembly AI	0	2	De-identified or aggregated only

3.6 Agreement between coders

Before adjudication the coders agreed on 77% to 98% of services per variable (Table 3). Agreement was highest on the opt-out mechanism and the naming of AI providers, and lowest on transcript retention and deletion time, where documents often give several periods for different data.

3.7 Tools named in institutional guidance

A companion survey of 130 universities recorded which tools institutions name as approved or prohibited for research transcription [1, 6]. Twelve of the tools they name are services in our frame. Table 4 sets the number of institutions that name each as approved or prohibited beside what the service's own documents say about training. The two analyses coded different documents and the institutional coding did not record why a tool was approved or barred, so the table is a juxtaposition, not a test of whether vendor terms explain institutional decisions.

The two platforms most often approved are coded differently: Zoom's documents say customer content is not used to train models, while Microsoft Teams' documents conflict. Of the three services

most often barred, Otter.ai's documents conflict, Read.ai's are unclear, and Fireflies.ai's say content is not used for training. Institutions that bar a service are therefore not, on this evidence, acting on a uniform reading of what its documents say about training. Because institutional guidance often names a product without giving its reasons, we cannot say whether training terms entered into any institution's decision.

4 Discussion

The main question is often not answered. Whether a recording is used to train models is the question users, institutions and regulators ask first. Only about a quarter of services answer it with a clear no, and for nearly half the documents are silent, unclear or contradictory. A user who reads only the privacy policy can come away with a different answer from one who reads the terms or the help center.

Contradictions are a disclosure problem in their own right. Seven services publish documents that disagree about training. Some contradictions probably reflect pages that were not updated when a policy changed; others reflect different terms for different plans. Either way, an ethics office that approves a tool on the strength of a help-center page may be relying on the document that does not govern. Keeping AI-use statements in one versioned document, or stating which document controls, would remove most of the conflicts we found.

Retention is described by purpose, not by time. Four in ten services keep audio "as long as necessary". That formula is common in privacy law and not wrong, but it does not let a researcher write a retention period into a data management plan. Services that state a fixed period, or delete audio after processing, give users something they can rely on.

Opt-outs are rarer than training. Where training is the default, about half the services describe no way to opt out, and some reserve opt-outs for business plans. Combined with the finding that business-tier terms differ at more than a quarter of services, this suggests that individual users and small teams receive weaker terms than enterprise customers.

Institutional decisions do not track the documents. Universities approve Zoom and bar Fireflies.ai and Read.ai, and the training terms in those services' own documents do not separate them cleanly (Section 3.7). Guidance that names a product without saying which document it relied on cannot be checked against the vendor's terms, and a vendor's contradictory documents leave an ethics office with no stable text to rely on.

The documents are becoming more explicit. Compared with 2024, many services now say something about training where they said nothing, and most of the new statements limit training. Litigation and institutional scrutiny may be part of the reason; this study cannot tell.

4.1 Limitations

The study reads documents, not systems; a service may do more or less than its documents say. Archive captures can lag the live page, and 38 documents were captured before April 2026. Some documents were JavaScript shells with no text, and one service had no retrievable documents. The frame is English-language and self-serve. Coding was language-model-assisted; every quote was machine-checked and every disagreement adjudicated against the text, but coders can still misread a passage, and the dataset lets readers check each code against its quote. The drift comparison is asymmetric, as described in Section 3.5. Named examples quote documents verbatim; they are not statements about any company's conduct.

5 Conclusion

Across 53 transcription and meeting-notes services, the public documents give a clear answer about training on customer recordings for about a quarter, permit default training at more than a quarter, and are silent, unclear or contradictory for nearly half. Retention is mostly open-ended and opt-outs are scarce where training is the default. Services could close most of these gaps by stating, in one controlling document, whether customer audio is used for training, how long it is kept, who can access it and which providers process it.

Statements

Data and materials availability. The coded dataset (final codes, both coders' codings, adjudication notes and verbatim quotes), document URLs, Wayback capture timestamps and archive URLs, SHA-256 hashes, the codebook, the frame, the plan files with their hash log, and the scripts are released under CC-BY-4.0 (code under MIT) on Zenodo (DOI: https://doi.org/10.5281/zenodo.24203123). Document texts are third-party material and are not redistributed; the archive URLs let anyone retrieve them. This version of the paper is identified as Preprint DOI: https://doi.org/10.5281/zenodo.24203179. The study was not registered in a public registry; the plan was hashed before data collection and the hashes are in the dataset.

Competing interests. The author is affiliated with Pepys, which develops and sells an AI transcription service. Pepys is one of the 53 services studied. It was coded by the same procedure as every other service, from captures of its public documents taken on 3 October 2026, and it appears in the same tables. The paper does not rank or recommend services. The author declares no other competing interests.

Vendor contact. The written plan said named vendors would receive a 14-day right of reply before publication. This was not done: no service was contacted before publication. Every vendor-specific statement rests on a verbatim quote, URL and capture date in the released dataset, and corrections are welcome at contact@pepys.co.

Funding. No external funding.

Use of AI tools. An AI system, Claude (Anthropic), was used under the author's direction for the following tasks: locating the public documents of each service by web search; both coding passes and the adjudication, under the written codebook, with automated checks that every quote appears verbatim in the saved document; writing and debugging the capture, agreement and analysis scripts that produce the tables; searching for and summarising the literature; and drafting and editing the manuscript text. The AI system is not an author and has no authorship role. The author reviewed the output, including the numbers, tables and references, and is responsible for the content.

Ethics. The study uses public documents only, retrieved from the Internet Archive. There are no human participants and no personal data. Services are quoted briefly and verbatim, with archive URLs, in the dataset.

Author contributions. Ankur Shrestha: conceptualization, design, data curation, analysis, writing, and supervision of AI-assisted work.

References

[1] A. Shrestha, "Institutional guidance on AI transcription for research interviews: A survey of 130 universities in the United States, United Kingdom and Canada," SSRN 7685681; preprint DOI 10.5281/zenodo.23846598, 2026. [Online]. Available: https://ssrn.com/abstract=7685681

[2] United States District Court, N.D. Cal., "In re Otter.AI Privacy Litigation, no. 5:25-cv-06911," 2026, order granting in part motion to dismiss, Aug. 13, 2026 (Judge Eumi K. Lee); allegations at the pleading stage.

[3] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves, K. Singh, and T. Xie, "PolicyLint: Investigating internal privacy policy contradictions on Google Play," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 585–602.

[4] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. Giovanni Leon, M. Schaarup Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1330–1340.

[5] J. King, K. Klyman, E. Capstick, T. Saade, and V. Hsieh, "User privacy and large language models: An analysis of frontier developers' privacy policies," arXiv:2509.05382, 2025. [Online]. Available: https://arxiv.org/abs/2509.05382

[6] A. Shrestha, "Institutional guidance on AI transcription for research interviews: Coded dataset (v1.0)," Zenodo dataset, 2026. [Online]. Available: https://doi.org/10.5281/zenodo.23846482

A Codebook summary

#	Variable	Values
1	Training on customer content	prohibited; opt-in; de-identified or aggregated only; yes by default; conflicting; ambiguous; not stated
2	Third-party providers training	prohibited; allowed; ambiguous; not stated
3	Opt-out mechanism	self-serve setting; request to vendor; enterprise only; none stated; not applicable
4	Audio retention	fixed period; until user deletes; as long as necessary; conflicting; not stated
5	Transcript retention	as for audio
6	Deletion after request	fixed period; unspecified; not stated
7	Human review	routine; conditional; never; not stated
8	Sub-processor disclosure	named list; categories only; none
9	AI providers named	named; categories only; not stated
10	Data location	list of regions, customer choice, or not stated
11	Biometric or voiceprint	mentions with consent terms; mentions without; disclaims; not stated
12	Business-tier differences	yes; none stated

Each value other than "not stated" carries at least one verbatim quote; the full codebook is in the dataset.

B Training code by service

Group	Prohibited	Opt-in	De-identified only	Yes, default	Conflicting	Unclear, not stated or unavailable
Consumer and prosumer (35)	Amberscript, Cleanvoice AI, GoTranscript, Good Tape, Kapwing, Pepys, Sonix, Transcribe by Wreally, Transkriptor, Trint, TurboScribe	Flixier, Happy Scribe	OpusClip, Zubtitle	TranscribeMe, VEED (free tier), Zeemo	Descript, Rev, Riverside, Vizard	Castmagic, Clipchamp, Cockatoo, MacWhisper, Maestra, Notta, Simon Says, Speech-Text.AI, Submagic, Swell AI, Temi, Vocalmatic, Vook.ai
Meeting-notes bots (10)	Fellow, Fireflies.ai	MeetGeek	Fathom, Sembly AI	–	Otter.ai	Avoma, Jamy AI, Laxis, Read AI
Quote-based (5)	–	Verbit Legal Visor	–	3Play Media	Athreon	Gong, Interprefy
Platforms (3)	Zoom	Google Meet	–	–	Microsoft Teams	–

Per-service codes for all twelve variables, with quotes and capture dates, are in the dataset.